

文書理解処理のための前処理について

2L-4

辻 正博、米田政明、長谷博行、酒井充
(富山大学工学部電子情報工学科)

1 はじめに

文書の自動読み取りを汎用化するためにはいくつかの入力形態を許容しなければならない。例えば、イメージスキャナーから読み取った画像として本の両ページを取り込んだ画像、片ページを取り込んだ画像、1枚の紙を読み取った画像などの形態が考えられる。従来、OCRなどの入力形態としては、対象とする画像を特定の形態に限定したり、認識領域を指定することによって、処理対象領域を求めている。

本研究では入力された画像から、書籍の置かれた位置を求め、それがどのような形態であるか判断し、処理対象であるページの領域を求めるものである。対象とする画像は、次の3種類である。

- 書籍の両ページを取り込んだ画像(見開き2ページ)
- 書籍の片ページを取り込んだ画像(片ページ)
- 1枚の紙を取り込んだ画像

実験では、見開きの文庫本、A5版の書籍、片ページのB5版の書籍、1枚の紙を読み込んだ130枚の画像に対して処理を行い127の画像に対して正しく処理することができた。

2 ページ抽出アルゴリズム

見開きの2ページを取り込んだ場合、左右のページでは、傾きが、必ずしも等しくないため、領域分割などの処理を、両ページまとめて行うことはできない。このような場合、領域分割や、文字認識を行う前に、2ページを分割し、それぞれのページを別々に処理しなければならない。領域理解を行った結果によって、対象画像の形態を判別する事も考えられるが、本研究では領域分割、文字認識の前処理として使用できる手法を提唱する。

2.1 処理対象の抽出

見開き境界の検出において、文書の縁の部分鮮明にわかる必要があるため、スキャナーの蓋を黒くして、文書の置いてない部分は、黒く写るようにしている。

このようにして、取り込まれた画像は、図1のようになる。はじめにこの入力画像から、処理対象となる文書の位置検出を行う。画像上に n 本のゾンデを左から右に走査させ最初に黒画素から白画素に変化する位置を記憶し、変化点とする。 n 個の変化点のうち最も密集している点群の平均をとり左の境界とする。同様に右、上、下方向からも走査を行い4つの境界を求め、処理対象範囲とする。

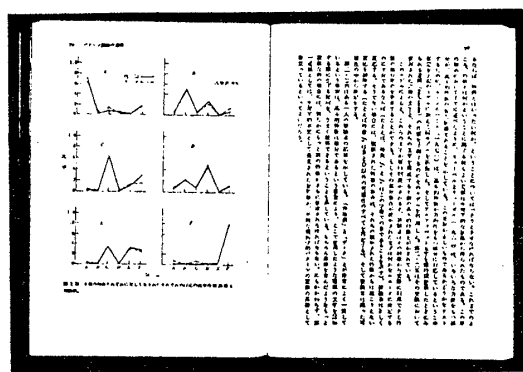


図1: 入力画像

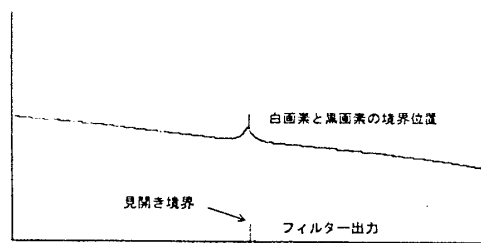


図2: 入力画像の下辺の状態

2.2 見開き境界位置の検出法

本を開いてスキャナーで読み取った場合、見開き境界の部分がガラス面から浮いているので黒く写る。本の縁の部分を見ると見開き境界のところは直線ではなく内側にへこんでいる。この変化に着目して、どのような入力形態の画像であるかの判定を行う。又、ページの見開き境界の両側には空白部分がある。この特徴を使って見開き境界の位置を求める。

見開きの画像の1辺の黒領域境界は図2のようになる。これより、大きく変化している部分を取り出すためにガウシアンラプラシアンフィルタをかけ、直流成分を取り除きかつ平滑化する。その結果を用いて次の1から5の手順により見開き境界を検出する。

1. フィルター出力が最大の位置を求める。高さが α を越えていればその位置をAとする。向かい合う辺についても求め、Bとする。
2. 一方の辺にしか α を越えるものがなければ、それを境界とする。
3. A、Bの位置が近接している場合、A、Bの中間を境界とする。
4. A、Bの位置が異なった場合
 - Aに向かい合う辺の対応する所に α を越える部分があればAを境界とする。
 - Bに向かい合う辺の対応する所に α を越える部分があればBを境界とする。
5. 求まった境界の位置の両側に幅 β 以上の空白領域があれば、見開き境界とする。

2. 3 入力形態の判定法

見開き境界の条件が成り立つ部分がどこにあるかによって、入力形態の判定を行う。処理対象がX方向よりY方向の方が長い縦長ならば、図3に示すような処理対象の左3分の1の領域をL、右3分の1の領域をR、Y方向の中間3分の1をCとする。次の条件によって入力形態の判定を行う。X方向の長さがY方向の長さより長い場合は、上記領域を90度回転させた領域を同様に考え、判定する。

- 見開き境界が領域L又はRに検出されれば片ページである。
- 見開き境界が領域Cに検出されれば見開きである。
- 見開き境界が検出されなければ1枚ものである。

3 実験と考察

本方式の有効性を確かめるためにイメージスキャナーにより130枚の画像を読み込み、入力形態の判定を行った。

データの内訳は、見開き2ページを取り込んだ文庫本20枚、A5版の本50枚、片ページを取り込んだB5版50枚、1枚ものの紙10枚である。入力画像はA4サイズ400dpiの2値画像として読み込み、Sun sparcl10によって処理を行った。

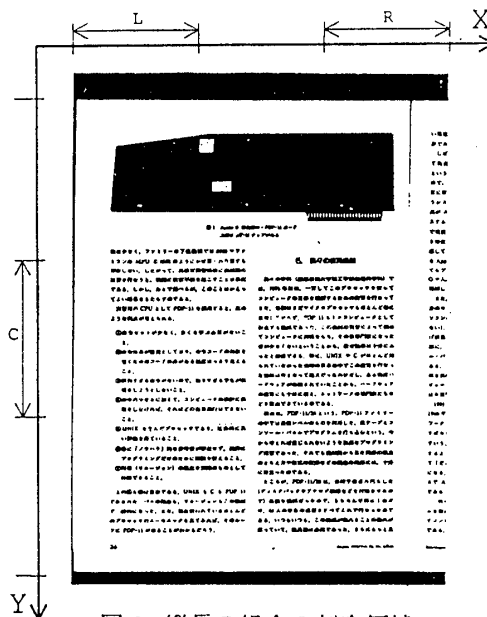


図3: 縦長の場合の判定領域

入力形態	資料数	正判定	誤判定
文庫本(見開き)	20	20	0
A5版(見開き)	50	49	1
B5版(片ページ)	50	48	2
1枚もの	10	10	0

表1: 実験結果

実験の結果を表1に示す。判定を誤ったもの3枚のうち1つは図4に示すように、ページの境界部分に空白が無かったために見開き境界の条件が成立しなかったものであった。他の2つは、境界部分がガラス面から浮かず、黒く写る部分が全くなかったものであった。

問題点として、判定を誤った3例のように、書籍によっては、アルゴリズムで仮定している見開き境界の条件が成り立たないものには、対応できないことが挙げられる。これを解決するには、更に他の条件を考慮していく必要があるが、すべてのものに対処することは必ずしも容易でない。本研究では利用形態の判定とページ領域を求めることを目的としたが、利用者に見開きであるのか、片ページなのかの入力をさせ、それを利用すれば今回の実験で誤判定した例に対しても正しく領域を求められるであろう。

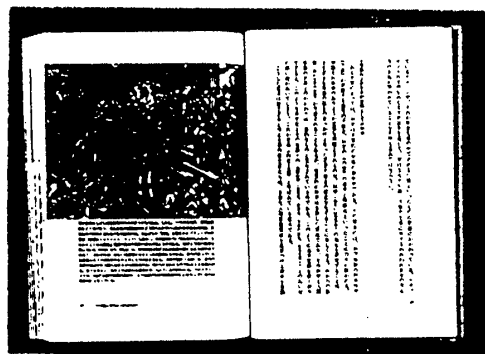


図4: 判定を誤った例